Title

Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO_2 concentrations.

Journal

Environmental microbiology, 19(2)

ISSN

1462-2912

Authors

Probst, Alexander J
Castelle, Cindy J
Singh, Andrea
et al.

Publication Date

2017-02-01

DOI

10.1111/1462-2920.13362

Peer reviewed

Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO_2 concentrations

Alexander J. Probst,[1] Cindy J. Castelle,[1] Andrea Singh,[1] Christopher T. Brown,[2] Karthik Anantharaman,[1] Itai Sharon,[1†] Laura A. Hug,[1†] David Burstein,[1] Joanne B. Emerson,[1§] Brian C. Thomas[1] and Jillian F. Banfield[1,3,4*]

[1]Department of Earth and Planetary Sciences, University of California, Berkeley, 307 McCone Hall, CA 94720, USA. [2]Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720, USA. [3]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA. [4]Earth Science Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA, USA.

*For correspondence. E-mail jbanfield@berkeley.edu; Tel. (+1) 510 316 4334; Fax (+1) 510 643 9980 Present addresses: † Migal – Galilee Research Institute, South Industrial Zone, Kiryat Shmona 11016, Israel, and Tel Hai College, M.P. Upper Galilee 12210, Israel; ‡ Department of Biology, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1; § Department of Microbiology, The Ohio State University, 496 West 12th Ave., Columbus, OH 43210, USA.

Summary

As in many deep underground environments, the microbial communities in subsurface high-CO_2 ecosystems remain relatively unexplored. Recent investigations based on single-gene assays revealed a remarkable variety of organisms from little studied phyla in Crystal Geyser (Utah, USA), a site where deeply sourced CO_2-saturated fluids are erupted at the surface. To provide genomic resolution of the metabolisms of these organisms, we used a novel metagenomic approach to recover 227 high-quality genomes from 150 microbial species affiliated with 46 different phylum-level lineages. Bacteria from two novel phylum-level lineages have the capacity for CO_2 fixation. Analyses of carbon fixation pathways in all studied organisms revealed that the Wood-Ljungdahl pathway and the Calvin-Benson-Bassham Cycle occurred with the highest frequency, whereas the reverse TCA cycle was little used. We infer that this, and selection for form II RuBisCOs, are adaptions to high CO_2-concentrations. However, many autotrophs can also grow mixotrophically, a strategy that confers metabolic versatility. The assignment of 156 hydrogenases to 90 different organisms suggests that H_2 is an important inter-species energy currency even under gaseous CO_2-saturation. Overall, metabolic analyses at the organism level provided insight into the biochemical cycles that support subsurface life under the extreme condition of CO_2 saturation.

Introduction

Earth's terrestrial subsurface is a massive reservoir of life and holds significance for human society, providing sources of groundwater, minerals and metals as well as oil and gas. Despite its fundamental scientific and industrial importance, the microbiology of the subsurface is only just beginning to come to light. Metagenomic studies of groundwater provided evidence that subsurface environments host a vast phylogenetic diversity of little known organisms (Nunoura *et al*., 2011; Brown *et al*., 2015; Castelle *et al*., 2015; Emerson *et al*., 2016), some with versatile metabolic potential (Castelle *et al*., 2013). More recently, a study of produced water samples from oil field reservoirs revealed the presence of multiple candidate phyla in environments more than 1000 m below the surface (Hu *et al*., 2016).

There is rising interest in the use of the subsurface for geological carbon sequestration, i.e. the capture and storage of anthropogenic CO_2 in the deep subsurface. The risks of geological carbon sequestration are many, and include the hazard of local leakage of CO_2 into near-surface environments. Such an event could affect potable water by mobilizing contaminants and interfere with subsurface groundwater ecosystems (Wilson *et al*., 2003). The consequences of considerable changes to the terrestrial biosphere are difficult to predict, but are likely significant given that these regions harbour a substantial fraction of life on Earth (Whitman *et al*., 1998), and their indigenous organisms play significant roles in biogeochemical cycling.

Restricted access to subsurface ecosystems has limited studies of the biological effects of high CO_2 concentrations. To date, marker gene specific surveys (based on 16S rRNA gene analysis) have indicated that high CO_2 concentrations alter the microbial composition of groundwater ecosystems (Wandrey *et al*., 2011a, 2011b; Mu *et al*., 2014). However, a weakness of this approach is that many organisms escape single-gene based approaches [due to primer mismatch, e.g. (Baker *et al*., 2006), and/or insertions that substantially increase gene length (Brown *et al*., 2015)]. Moreover, these assays do not allow inference of metabolic potential of organisms in an ecosystem. These limitations may be circumvented using genome-resolved metagenomic methods to determine community composition and metabolic adaption of autotrophs, i.e. by which pathway they capture CO_2 into organic molecules (CO_2 fixation). For example, the genomes of iron and sulfur oxidizing bacteria reconstructed from aquifer sediment communities encode forms I and/or II RuBisCOs of the Calvin-Benson-Bassham (CBB) cycle (Handley *et al*., 2013). Further, assays of single gene RuBisCO transcripts from a pristine limestone aquifer provided information about the activity of the CBB pathway (Herrmann *et al*., 2015). Genomic insights also revealed that uncultivated sediment-associated Chloroflexi use the Wood Ljungdahl (WL) pathway for CO_2 fixation, a pathway not previously known in this phylum (Hug *et al*., 2013). A novel version of the WL pathway was identified in *Candidatus* ‘Altiarchaeum hamiconexum’ SM1 from a cold sulfide spring and a cold CO_2-driven geyser (Probst *et al*., 2014).

Sites of discharge of deeply sourced groundwater enable sampling of subsurface microbial communities with minimal contamination and without requirement for drilling. Crystal Geyser in Utah, USA, is a site where CO_2-saturated fluids sourced from depths of up to 800 m erupt to the surface via a well installed in 1936 for hydrocarbon exploration (Baer and Rigby, 1978). The geyser is widely considered to be an analogue for a carbon sequestration leakage site (Shipton *et al*., 2006; Friedmann, 2007; Lewicki *et al*., 2007; Bickle, 2009), and many hydrogeological and geochemical studies have been conducted at this site. These studies have thoroughly described the periodic eruptions of the geyser (Murray, 1989; Gouveia and Friedmann, 2006; Bryan, 2008; Han *et al*., 2013; Watson *et al*., 2014) and its water and gas composition (Mayo *et al*., 1991; Evans *et al*., 2004; Shipton *et al*., 2006; Heath *et al*., 2009). CO_2 was determined to be the only gas component in the geyser, as traces of other gases like O_2 are probably contaminants as they occur at atmospheric ratios in some measurements (Mayo *et al*., 1991; Heath *et al*., 2009). Although the source of the CO_2 in the geyser has not yet been fully identified, recent data suggests that decarbonation of carbonates and clay-carbonate reactions involving rocks at ~800 m depth are responsible for the high CO_2 discharge (Heath *et al*., 2009). These features make Crystal Geyser an ideal site for studying subsurface ecosystems in CO_2-saturated environments and for determining the metabolic potential of the resident organisms.

Prior phylogenetic investigation at the Crystal Geyser site using ribosomal protein S3 and 16S rRNA genes indicated a great diversity of previously unstudied or little studied archaea and bacteria (Emerson *et al*., 2016). The authors also reported a large number of RuBisCO genes implicated to function in carbon fixation via the CBB cycle (other pathways for CO_2 fixation were not considered). The existing genomic resolution involves only four near complete and six partial genomes, and the majority of these genomes were from relatively well-studied lineages. Thus, the prior analyses were insufficient for analysis of the prevalence of different carbon fixation pathways and prediction of carbon, nitrogen or sulphur cycles within the community. Here, we utilized information from eleven metagenomes sequenced from water filtered over two eruptions of the Crystal Geyser and applied a novel strategy of genome recovery that integrated three different binning algorithms to obtain genomes for organisms that span a substantial fraction of the phylogenomic diversity present. This enabled us to reconstruct metabolic pathways for a large fraction of the organisms in the community. Using this information, we tested the prediction that CO_2 exerts a selective pressure on the pathways by which CO_2 fixation occurs. Further, we investigated the possible energy resources that support autotrophy and evaluated potential ways in which heterotrophic community members impact the metabolism of autotrophs.

Results and discussion

Integration of different binning algorithms enabled high genomic resolution of the stable community in the Crystal Geyser system

A variety of different microorganisms have been detected in high CO_2 subsurface aquifers (Emerson *et al*., 2016; Herrmann *et al*., 2015; Santillan *et al*., 2015) but very little is known about the metabolisms of these organisms, particularly about their genetic adaptations to high-CO_2 concentrations and their nutrient cycles. Here, we predicted the metabolic capacities of organisms in the high-CO_2 Crystal Geyser ecosystem based on cultivation-independent reconstruction of microbial genomes. We collected samples from two CO_2-driven geyser eruptions. Fluid samples from the first eruption were filtered onto eight replicate 3-µm filters (Supporting Information Table S-1). Filters were changed after they clogged, which occurred at different time points, likely resulting in different cell size distributions across the samples. Samples from another eruption event that occurred 32 h later were sequentially collected onto a 3-µm pore size filter (CG2_3.0) and a 0.2-µm pore size filter (CG1_0.2). Analysis of a first small increment of sequence data from the CG1_0.2 sample yielded three near-complete and seven partial microbial genomes (sample CGorig, entire sample set and metadata presented in Supporting Information Table S-1) (Emerson *et al*., 2016). To put this into perspective, ribosomal protein S3 sequence profiling of the full CG1_0.2 and CG2_3.0 samples indicated the presence of around 300 distinct genotypes (Emerson *et al*., 2016).

Using data from the eleven different metagenomic datasets we binned 227 high-quality microbial genomes, i.e. genomes that underwent manual curation regarding CG-content, coverage, and phylogeny of each single scaffold and that had a completeness >70% based on 51 bacterial and 38 archaeal single-copy genes (Supporting Information Fig. S-1). These genomes correspond to 150 different bacterial and archaeal species present in the Crystal Geyser system (Fig. 1, completeness of each genome provided in Supporting Information Fig. S-1). Critical to the success of the genome recovery effort was the sequential application of binning methods that made use of sequence composition or sequence composition coupled to abundance data (work flow provided in Supporting Information Fig. S-2, for details on the approach see experimental procedures). Binning efforts focused on the metagenomes CG1_0.2 and CG2_3.0, as these had the greatest sequencing depth (binning results in Supporting Information Fig. S-3). Since samples CG1_0.2 and CG2_3.0 came from the same eruption event, we de-replicated the 227 genomes from both samples by whole genome alignment and clustering at >98% nucleotide identity to retrieve one genome per sampled species (150 in total). This approach provided significantly more near-complete microbial genomes than each single binning approach (Supporting Information Fig. S-2-B).

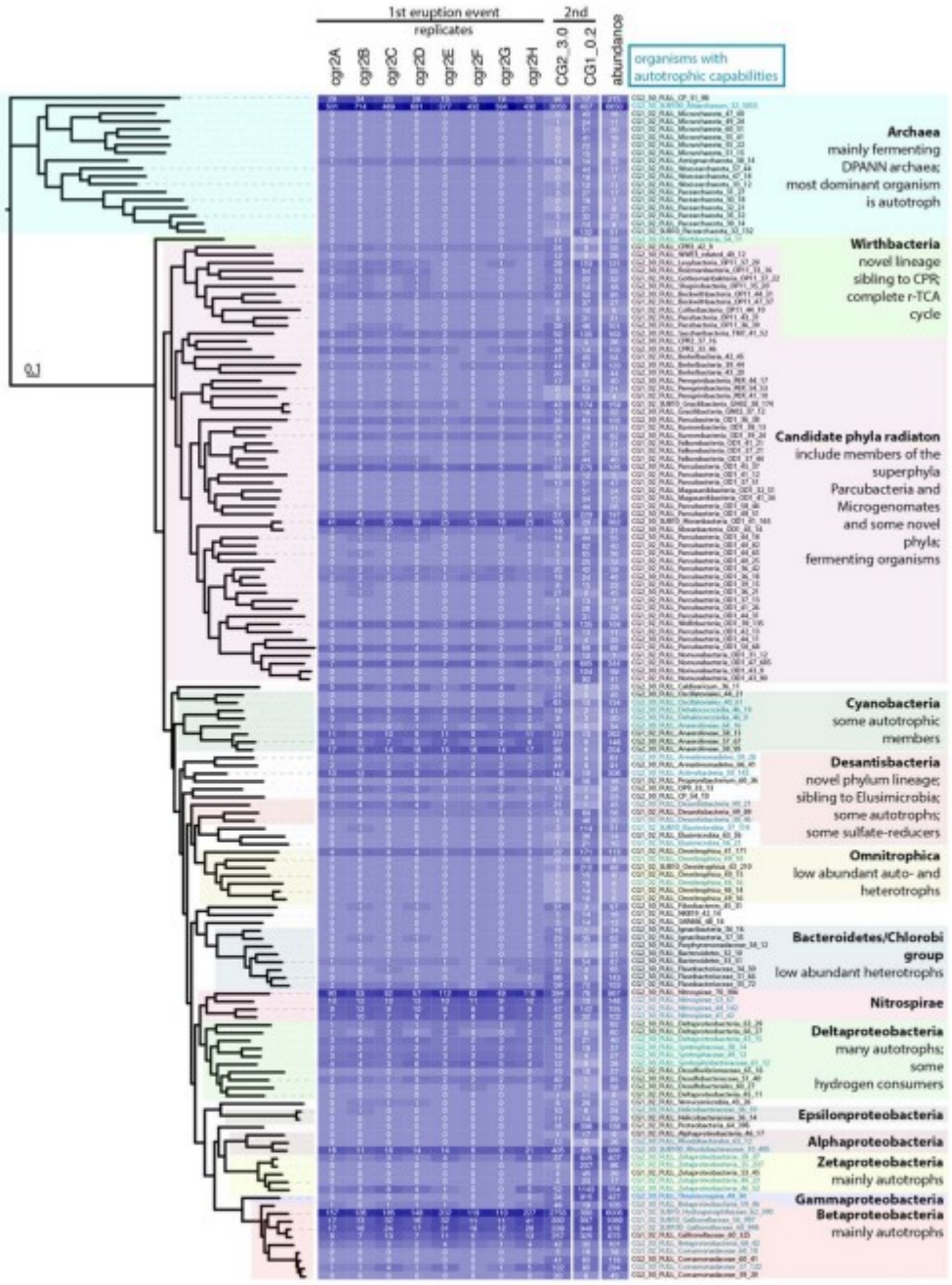

Fig. 1. Overview of genomes recovered from Crystal Geyser and their relative abundance. Heatmap is based on log10 abundances (z-normalized by column) of each genome in the different samples. Relative abundance values (given within each cell of the heatmap) are based on stringent read mapping and are corrected for genome length. Relative abundances of genomes across samples from the first eruption event (cgrA-H) show high correlation with abundances from sample CG2_3.0 of the second eruption event. Please note that samples cgrA-H have less sequencing coverage than sample CG2_3.0, resulting in absence of low abundant members. Abundance values in the very right column are relative abundances calculated from read mapping of sample CG1_0.2 and CG2_3.0 taking into account the amount of DNA extracted and the amount of volume filtered (Table S1). Phylogenetic tree was reconstructed using 16 concatenated ribosomal proteins, members with autotrophic capabilities are highlighted with blue font. For detailed phylogenetic placement of each lineage within the tree of life please see Supporting Information Files 1 and 2.

The eight samples collected onto the 3-µm filters (cgr2A-H) and the second 3-µm filter sample (CG2_3.0) all had very similar community compositions regarding the organisms detected in these samples (Fig. 1). A Pearson correlation of relative abundances of all organisms present in CG2_3.0

against all those of 3.0-µm metagenomes from the first eruption event (cgr2A-H) provided evidence that the community compositions of the two sampling events are highly similar (p-value $<10^{-15}$, correlation >0.84). Despite this, the organism abundance patterns were different enough over the sample series to provide binning power for differential-coverage ESOMs and ABAWACA.

The recovered 150 genomes represent the majority microbial community members present at >~0.1% of the communities sampled in the 0.2-µm and 3.0-µm filter datasets CG1_0.2 and CG2_3.0 (based on rpS3 inventories). Genomes were reconstructed for 41 and 38 of the 50 most abundant organisms sampled in CG1_0.2 and CG2_3.0 respectively (Supporting Information Fig. S-4). Thus, the collection of genomes enabled us to achieve a relatively comprehensive understanding of the metabolic potential of a significant fraction of the stable community in the ecosystem (relative change of each organism across the metagenomic samples is reported in Supporting Information Fig. S-5).

An important objective of this study was to investigate the nutrient cycling of the microbial community including metabolic pathways by which CO_2 fixation can occur and estimate their frequency in the microbial community. Confident metabolic prediction could only be done for microorganisms for which we reconstructed high-quality genomes (>70% completeness, no detectable contamination). Thus, subsequent analyses estimate the frequency of pathways of interest in the fractions of these communities that are represented by these 150 genomes.

150 genomes for a diversity of microbes, including bacteria from novel phylum-level lineages

Based on phylogenetic analysis of concatenated ribosomal protein sequences, the recovered genomes represent organisms from at least 42 different microbial phyla and four potentially novel phylum-level lineages for which no prior genomic sampling exists (Fig. 1, detailed phylogenetic trees in Supporting Information File1 and File2). Most highly abundant genomes were two previously described autotrophs, the archaeon '*Ca*. Altiarchaeum' (Probst *et al*., 2014), followed by a member of the betaproteobacterial family Hydrogenophilaceae (Fig. 1, rank abundance curve in Supporting Information Fig. S-6A) (Emerson *et al*., 2016).

More than one third of the recovered genomes (56) are from bacteria that are phylogenetically affiliated with organisms with limited metabolic capabilities from a monophyletic group of phyla and referred to as the candidate phyla radiation [CPR (Brown *et al*., 2015)]. As reported previously, some carry inserts in their 16S rRNA genes. These insertions, as well as unusual 16S rRNA gene sequences, would have rendered some of these organisms invisible in standard 16S rRNA gene-based surveys, as noted previously (Brown *et al*., 2015). Interestingly, two of the nine

Microgenomates (OP11) have ribosomal protein L9, also predicted to be rare in Microgenomates from another ecosystem (Brown *et al*., 2015).

Sixteen genomes belonged to members of the DPANN Archaea, a group that has been previously studied in groundwater (Castelle *et al*., 2015). Many predicted proteins in the CPR bacterial and DPANN archaeal genomes cannot be annotated to date – a common trait of these lineages across ecosystems (ordination analysis of metabolic profiles in Supporting Information Fig. S-6C) (Wrighton *et al*., 2012; Brown *et al*., 2015; Castelle *et al*., 2015).

Also detected within the autotrophic community were representatives of the candidate phylum Omnitrophica (Rinke *et al*., 2013), previously known as OP3 (Hugenholtz *et al*., 1998; Kolinko *et al*., 2012). In total seven genomes for members of this phylum were recovered, with an estimated completeness of up to 92%. While three of the members had group 4 membrane-bound hydrogenases (hydrogen evolving) (Vignais and Colbeau, 2004) indicative of fermentative metabolism (see below), two had autotrophic capabilities coupled to sulfate reduction. Omnitrophica appeared to be of low relative abundance compared to other members of the community (Fig. 1) and thus cannot be seen as one of the most important primary producers in the Crystal Geyser system. However, this candidate phylum appeared to be widespread across diverse ecosystems and its members potentially function as global primary producers in the subsurface (Glöckner *et al*., 2010; Kolinko *et al*., 2012; Rinke *et al*., 2013; Kolinko *et al*., 2015). Here we show that these bacteria can exist under extremely high CO_2 conditions, as found at Crystal Geyser.

We identified organisms that belong to a previously undetected phylum-level lineage based on their distinct phylogenetic position within the concatenated ribosomal protein tree of bacteria (Supporting Information File 1). We propose the name ‘Desantisbacteria’ for this potential candidate phylum. This name honors Todd Z. DeSantis, who performed pioneering work in 16S rRNA gene research (Greengenes, NAST). The clade is represented by three genomes (CG2_30_FULL_Desantisbacteria_40_21, CG1_02_FULL_Desantisbacteria_38_46, and CG1_02_FULL_Desantisbacteria_49_89) and places sibling to Elusimicrobia (Fig. 2A). This phylogenetic placement was confirmed using 16S rRNA gene sequences which showed an average distance of 24% to the next classified sequence in SILVA (Pruesse *et al*., 2007; Yarza *et al*., 2014).

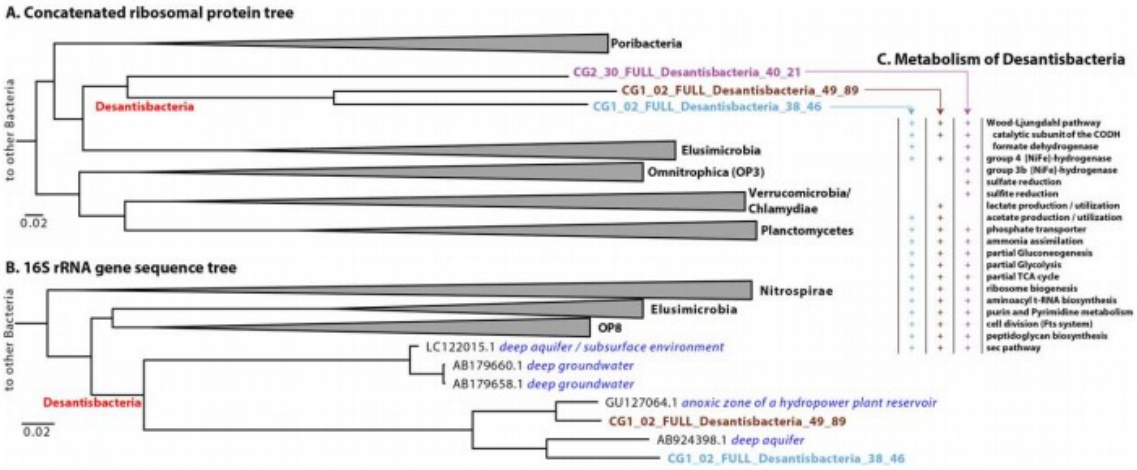

Fig. 2. Phylogeny and metabolism of *Ca.* 'Desantisbacteria'. A. Phylogenetic tree based on 16 concatenated ribosomal proteins reveals the distinct grouping of *Ca.* 'Desantisbacteria' genomes in the tree of life. B. Phylogeny based on 16S rRNA genes places the *Ca.* 'Desantisbacteria' near the Elusimicrobia and OP8. 16S rRNA gene sequences closely related to those from *Ca.* 'Desantisbacteria' were recovered from deep subsurface or freshwater ecosystems (dark blue). C. Comparative genomics of the three *Ca.* 'Desantisbacteria' genomes. A detailed list of shared KOs can be found in Supporting Information Table S-2.

The phylogenetic analyses also identified a new highly divergent lineage that branches as a sibling lineage to the CPR in the concatenated ribosomal protein tree (Fig. 3A). Searches of public databases with marker genes from this genome, including rpS3, have not yielded any significant hits in other datasets, emphasizing the novelty of this organism but making its phylogenetic placement very difficult. Unfortunately, the 16S rRNA gene was not identified, precluding comparison of this organism with organisms detected in cultivation independent gene surveys. Based on the extent of divergence in the concatenated ribosomal protein tree and the availability of a near-complete genome (see below), we propose the name species *Ca.* 'Wirthibacter wanneri' within the candidate phylum 'Wirthbacteria'. This name recognizes the scientific accomplishments of Prof. Dr. Reinhard Wirth, University of Regensburg, Germany, in the field of chemolithoautotrophs and their swimming behaviour, and Prof. Dr. Gerhard Wanner, University of Munich, Germany, in the field of ultrastructural analyses of microorganisms.

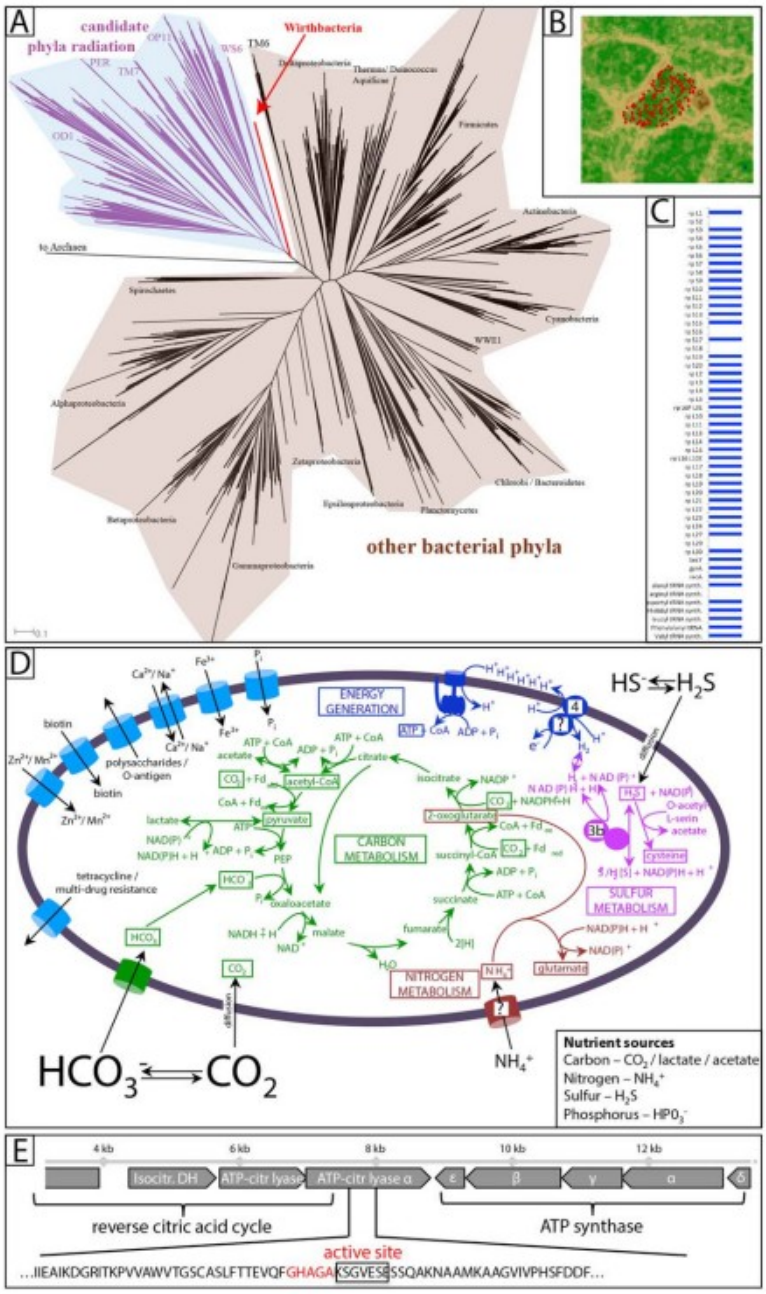

Fig. 3. Genomic analysis of *Ca.* 'Wirthibacter wanneri'. A. Phylogenetic placement of *Ca.* 'Wirthibacter wanneri' in the bacterial tree of life shows its close association with the CPR. B. ESOM map based on tetranucleotide frequencies of sample CG2_3.0. The *Ca.* 'Wirthibacter wanneri' form a distinct cluster. C. Genome completeness estimation of *Ca.* 'Wirthibacter wanneri' based on 51 bacterial single copy genes. D. Metabolic overview of important biochemical reactions for *Ca.* 'Wirthibacter wanneri' predicted from genomic evidence. E. Operon of ATP-citrate lyase and other genes associated with the r-TCA. These are located next to the ATP synthase encoding genes in *Ca.* 'Wirthibacter wanneri'. Notably, the active site of the ATP-citrate lyase subunit A is conserved (framed amino acids).

Candidatus 'Desantisbacteria' – novel anaerobic bacteria from diverse subsurface ecosystems

Together with 16S rRNA gene sequences available in public databases [NCBI and SILVA, (Pruesse *et al*., 2007)], the novel phylum-level lineage *Ca*. 'Desantisbacteria' was detected in five different ecosystems, all of which are in the subsurface and mostly reported to be anoxic according to Genbank entries (Fig 2B). This finding is in accordance with a strictly anaerobic metabolism, as indicated by O_2-sensitive pyruvate synthases in all three genomes. The genomes shared core metabolic pathways including nucleotide and peptidoglycan biosynthesis as well as phosphate and

ammonia assimilation via corresponding transporters (shared KEGG orthologies are presented in Supporting Information Table S-2).

The carbon metabolism appeared to rely on the WL pathway with the carbonmonoxide dehydrogenase / acetyl-CoA synthase (CO-DH/ACS) complex as its core. However, only two genomes encoded formate dehydrogenase, a key enzyme for carbon fixation via the WL pathway and were thus inferred to be autotrophs. Absence of this essential enzyme may be attributed to genome incompleteness. Acetate utilization/production was restricted to the two more closely related organisms (Fig 2A). CG1_02_FULL_Desantisbacteria_49_89, was inferred to make additional use of lactate. While all three organisms encoded a group 4 hydrogenase, which is usually used to produce H_2 (see below and Table 1), CG2_30_FULL_Desantisbacteria_40_21 had additionally a group 3b hydrogenase. The combination of group 3b hydrogenases, which can be utilized for H_2 consumption, and with the capacity for sulfate and sulfite reduction, may enable CG2_30_FULL_Desantisbacteria_40_21 to grow as a chemolithoautotroph.

Table 1. Classification of the [NiFe]-hydrogenases following the method of Vignais and Colbeau revealed that hydrogenases from the Crystal Geyser ecosystem are diverse, clustering into seven different groups: Group 1, 2a, 2b, 3b, 3c, 3d and 4 with different tendencies regarding hydrogen production or consumption (Vignais and Colbeau, 2004).

Type	Group	Number of times detected	Number of organisms	Relative abundance of organisms	Potential function in community
[NiFe]	Group 1	27	21	38.2%	H_2 consumption
	Group 2a	3	2	0.5%	H_2 consumption
	Group 2b	4	4	2.6%	H_2 sensing
	Group 3b	32	26	6.5%	H_2/H_2S production or consumption
	Group 3c	3	3	0.4%	unclear
	Group 3d	17	15	27.8%	H_2 consumption
	Group 4	55	47	95.7%	H_2 production/ferredoxin: NADP+ oxidoreductases
[FeFe]		15	15	4.1%	H_2 production

Detailed metabolic analysis of Candidatus ‘Wirthibacter wanneri’

The CPR lineage (OD1, OP11), which has many representative phyla, is composed of organisms that share limited metabolic capacities based the recovered genomic information (Brown *et al*., 2015). With incomplete pathways for nucleotide and lipid biosynthesis, these organisms were inferred to be heterotrophs and some have already been shown to live as symbionts (Gong *et al*., 2014; He *et al*., 2015) or need very rich media to be cultured (Soro *et al*., 2014).

Overall, the metabolic characteristics of *Ca*. ‘W. wanneri’ share some similarity with those reported for CPR. Given this, and the tentative phylogenetic placement noted above, we compare and contrast its predicted capacities to those predicted for the CPR. As for CPR bacteria, the genome is small, 1.46 Mbps. It encodes 33 tRNAs and 1,426 coding sequences with an average length of 924 bps. The genome encodes ribosomal protein L30,

unlike those of CPR bacteria. The GC content was 54.08% and the genome completeness was estimated to be 90%, based on inventory of single copy genes (Fig. 3C).

The near-complete *Ca*. 'W. wanneri' genome encodes a partial oxidative TCA cycle (o-TCA) and the full r-TCA cycle. Thus, we conclude that *Ca*. 'W. wanneri' can fix CO_2 via a complete r-TCA cycle. This capacity distinguishes *Ca*. 'W. wanneri' from CPR, as autotrophy (other than possibly via a RuBisCO type II/III–based pathway) has not been proposed for CPR bacteria (Sato *et al*., 2007; Wrighton *et al*., 2012; Kantor *et al*., 2013). The carbon species used for fixation via the r-TCA cycle are CO_2 and bicarbonate, and a trans-membrane bicarbonate transporter is encoded in the genome (Fig. 3D). The r-TCA cycle provides pyruvate and acetyl-CoA for biosynthesis. The presence of a lactate dehydrogenase indicates the ability to make lactate from pyruvate or to convert lactate to pyruvate for use in the in r-TCA cycle. Similarly, acetyl-CoA could be formed from acetate via an acetyl-CoA synthetase and enter the r-TCA as another form of mixotrophy.

Wirthbacteria pyrimidine and purine biosynthesis pathways, including production of the D-5-Phospho-alpha-D-ribose 1-diphosphate via the pentose phosphate pathway, were largely complete. Recent analysis of complete genome sequences of Peregrinibacteria CPR also identified these pathways (Anantharaman *et al*., 2016), although complete nucleotide biosynthesis pathways are likely absent in most other CPR (Brown *et al*., 2015).

While carbon, sulphur, nitrogen and phosphorus sources of *Ca*. 'W. wanneri' could be deduced from the genomic information (Fig. 3D), its energy metabolism was only partially resolved. Like the CPR, *Ca*. 'W. wanneri' lacks most of electron transport chain complexes and also terminal oxidases and reductases. Genes for an ATP-synthase were identified, thus the organism can use a proton gradient for energy generation. In fact, these genes were located right next to those encoding the reverse TCA cycle suggesting that carbon fixation and energy generation are closely intertwined (Fig. 3E). However, the mechanism for establishing the proton gradient remains unidentified. Although a group 4 hydrogenase could translocate protons across the membrane (Marreiros *et al*., 2013), the source of electrons (e.g. from formate dehydrogenase or a CO-DH) is unclear. Ferredoxin could be the reductant, but the source of reduced ferredoxin is uncertain (reduced ferredoxin is also required at two steps in the r-TCA cycle).

Ca. 'W. wanneri' is predicted to have a fermentative metabolism, based on the presence of lactate dehydrogenase (Wrighton *et al*., 2012; Kantor *et al*., 2013; Wrighton *et al*., 2014; Brown *et al*., 2015) as well as hydrogenases 3b and 4-types. Thus, given the inferred capacity to fix CO_2 and also to assimilate acetate, we conclude that this organism is a facultative autotroph or mixotroph. An intriguing possibility is that the *Ca*. 'W. wanneri' may be an intermediate between the CPR and other bacteria, both from the perspective

of its evolution (phylogenetic placement) and metabolism (due to the presence of the complete r-TCA cycle).

Prediction of how carbon, nitrogen, sulfur and hydrogen cycles support the autotrophic community

Of the 150 species' genomes analyzed, 38 (25%) are predicted to encode autotrophic capabilities (Fig. 1, Supporting Information Fig. S-6). Taxa with autotrophic capabilities span at least 13 different bacterial and archaeal phyla (Fig. 1).

In order to determine the representation of genomically characterized organisms capable of carbon fixation in the community, we utilized relative abundance measures based on stringent read mapping to the 150 genomes. Based on this analysis, these autotrophs accounted for >70% of the community (Supporting Information Fig. S-6B). As noted above, two organisms capable of CO_2 fixation are the most abundant organisms in the ecosystem. Not included in these statistics are (i) a member of the Cyanobacteria phylum (CG2_30_FULL_ Oscillatoriales_44_21, relative abundance ~0.2%) that is missing only the catalytic subunit of RuBisCO from the CBB cycle (probably due to genome fragmentation); (ii) One representative of the genus *Caldisericum* (CG2_30_FULL_Caldisericum_36_11), which has the two key enzymes for the reverse tricarboxylic acid cycle (r-TCA) but is missing other parts of this cycle and (iii) many organisms with RuBisCO form II/III and form III genes (Wrighton *et al*., 2012). These genes are suggested to function in the AMP salvage pathway and may fix one CO_2 per recycled AMP (Supporting Information Fig. S-8) (Sato *et al*., 2007). However, it is unknown if this is sufficient to support a purely autotrophic lifestyle.

Beyond carbon fixation, the capability of producing and utilizing lactate, acetate and formate were common traits in both autotrophs and heterotrophs, detected in 54%, 65% and 63% by relative abundance of the studied organisms respectively (Fig. 4, a detailed list of metabolic genes and pathways of all organisms is provided in Supporting Information Fig. S-7). The extent of formate utilization in autotrophs that use the WL pathway is difficult to separate from their autotrophic capabilities, as the critical enzyme (formate dehydrogenase) is required in both pathways. However, genes specific for acetate and lactate utilization in autotrophs may indicate the capacity for facultative autotrophy. We conclude that short-chain carboxylic acids (acetate, lactate, and formate) – probably produced by fermentative organisms from the abundant primary producers' cell remnants – are important substrates in carbon cycling in this high-CO_2 environment. Most of the autotrophs in the system can recycle this reduced form of organic carbon, likely to save energy by avoiding carbon fixation. Thus, most autotrophic members are likely to be mixotrophs similar to *Ca*. 'W. wanneri'.

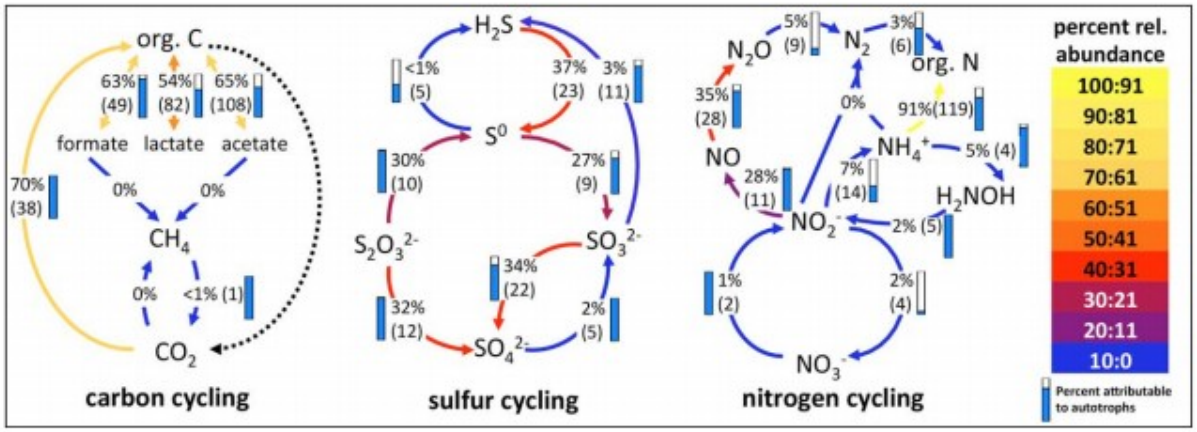

Fig. 4. Metabolic cycles predicted from 150 genomes reconstructed from the Crystal Geyser eruption fluid. Predicted carbon cycling shows a high dependency of the community on short-chain carboxylic acids (formate, lactate and acetate). Sulphur cycling revealed thiosulfate, sulphite and hydrogen sulphide as key metabolic substrates for the community. N cycling is dominated by conversion of nitrite to nitrous oxide through nitric oxide and ammonia assimilation is the most prevalent pathway to acquire nitrogen for molecular building blocks. Autotrophs dominate most of these processes; % values indicate the relative abundance of the organisms that can carry out each function, numbers indicate the number of organisms with that function and bar graphs are based on relative abundances of the function as contributed by autotrophs, calculated as shown in Fig. 1. Detailed metabolic analysis can be found in Supporting Information Fig. S-7.

Nitrogen and sulfur cycling can mostly be attributed to autotrophs in the system (bar charts in Fig. 4). The oxidation of (poly)sulfide, the oxidation of sulfite to sulfate, and the disproportionation of thiosulfate were prevalent capacities in community members (Fig. 4). Sulfate is present in Crystal Geyser fluids (at ~2000 ppm [Emerson *et al*., 2016]), as is sulfide (gas phase ~2 ppm). At pH ~6 of the geyser water (Emerson *et al*., 2016), sulfide is present as dissolved HS^-. The source of HS^- is unclear, as only five low abundance autotrophs showed the capacity to convert sulfate to sulfide. However, the genes for non-thermophilic sulfate reduction by archaea are unknown, so it is possible that '*Ca*. Altiarchaeum' is an anaerobic sulfate reducer (Probst *et al*., 2014).

Nitrate, the preferred electron acceptor under anaerobic conditions, is present at only trace levels in Crystal Geyser fluids (Emerson *et al*., 2016). Consistent with this, nitrate reductases or nitrite oxidoreductases were detected in only eleven relatively low abundance organisms. However, seven of these organisms have genes to convert nitrite to N_2O through NO and three organisms could reduce nitrite fully to N_2. Notably, 29 organisms (35% of the studied community) have the capacity to reduce NO to N_2O. Three moderately abundant organisms can fix N_2, and 120 out of 150 members (91% of the community in relative abundance) can assimilate ammonia (for details on predicted metabolic capacities see Supporting Information Fig. S-7).

H_2, the compound with the lowest redox potential in any ecosystem, could serve as a major energy substrate, potentially supporting CO_2 fixation in the Crystal Geyser ecosystem. H_2 has not been detected in the geyser fluids (Mayo *et al*., 1991; Heath *et al*., 2009), but this does not rule out its role as an energy currency because it may be rapidly cycled and genome analyses

provide evidence pointing to significant hydrogen metabolism. Analyses revealed that 90 of the 150 genomically characterized organisms are predicted to use and/or produce H_2 (Table 1, details in Supporting Information Table S-3). We identified 15 [FeFe]-hydrogenases based on three genes of confurcating complexes within the reconstructed genomes. These hydrogenases are commonly found in fermenters known to produce high molar ratios of H_2 (Schut and Adams, 2009; Sieber *et al*., 2012). More commonly, we identified [NiFe]-hydrogenases, which are implicated in both H_2 production and utilization (Table 1). Thus we infer the presence of an active hydrogen cycle in the ecosystem.

Group 2a [NiFe]-hydrogenases have been proposed to provide reductant to the r-TCA cycle for CO_2 fixation (Vignais and Colbeau, 2004). Given that the genome of a *Syntrophobacterales* in the Crystal Geyser system has the WL but not the r-TCA cycle, we conclude that the group 2a [NiFe]-hydrogenases encoded in its genome may be used for chemolithoautotrophic growth via the WL pathway, likely coupled to sulphate reduction (Fig. 5). (A detailed discussion about the type of hydrogenases and their potential functions beyond the summary of Table 1 are provided in the Supporting Information.)

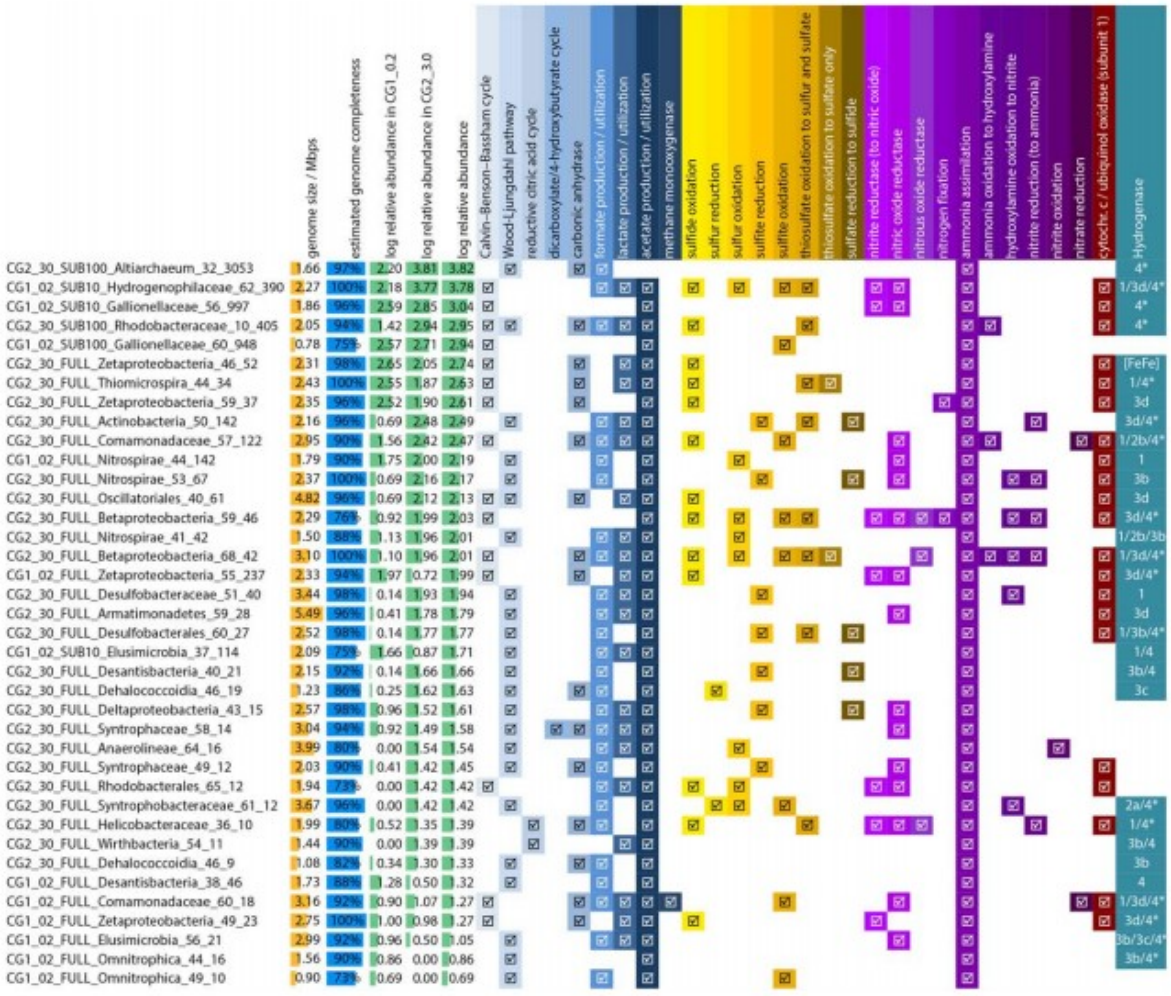

Fig. 5. Features of organisms with autotrophic potential from Crystal Geyser. Genome completeness was estimated from single copy genes (Supporting Information Fig. S-1). Displayed metabolic processes are according to Supporting Information Fig S-7 and Fig. 4. Organisms are ordered in decreasing relative abundance. Types of [NiFe] hydrogenases are indicated by designation to the corresponding group. *Catalytic residue not conserved in group 4 [NiFe] hydrogenase. For details regarding the groups of hydrogenases and their functions please see Supporting Information and Table 1.

As in some non-high-CO_2 environments (Greening *et al*., 2015), the presence of a great diversity of hydrogenases suggests that hydrogen is an important energy resource in this ecosystem. In fact, oxidation of H_2 derived from fermentation and geologic sources may largely support organisms responsible for CO_2 fixation in the Crystal Geyser ecosystem.

In order to evaluate the potential of microbial communities to impact CO_2 sequestration outcomes, we connected traits of interest to elements that would be supplied by leakage of geologically sequestered CO_2 and associated compounds. Leakage of CO_2 could mobilize metals (White *et al*., 2003) such as Fe(II) (Wilson *et al*., 2003), which could be immobilized by the Fe-oxidizing microorganisms present in the system (Emerson *et al*., 2016). The purity of the CO_2 stream used for geological carbon sequestration varies depending on its source (Porter *et al*., 2015). Oxidation products from coal or biomass generate impurities such as CO, H_2S, NH_3, NO_x, H_2 and CH_4 (Porter *et*

al., 2015). Our genomic analyses suggest that all of these components could support autotrophic growth of microorganisms indigenous to high-CO_2 environments. In fact, the oxidation of H_2S and the reduction of NO_x compounds can mainly be attributed to autotrophic members of the community (bar charts in Fig. 4). CO is an energy rich compound that can be used via the WL pathway directly for carbon fixation and NH_3 was shown to likely be the most important nitrogen source for organisms studied herein.

High CO_2 concentrations select for two main carbon fixation pathways

It has been reported in the literature that different carbon fixation pathways are adaptions to ecological niches (Berg, 2011). Pathways by which carbon fixation occurs can be a function of carbon dioxide concentrations (Markert *et al*., 2007; Berg, 2011) or a function of available energy (Könneke *et al*., 2014). However, little is known about how saturated CO_2 concentrations in a subsurface ecosystem influence the CO_2 fixation pathways and the frequency with which they occur in microorganisms. Thus, we evaluated the variety and frequency of the six known carbon-fixation pathways represented in the 150 genomically sampled organisms in the high CO_2 Crystal Geyser ecosystem (Fig. 5).

We detected a selection for specific forms of RuBisCO used in the CBB pathway. Among all organisms with the CBB cycle, only one encoded a form IB, while all others had form II (tree of RuBisCO genes is given in Supporting Information Fig. S-8). Form II RuBisCO was shown to perform best under high CO_2 concentrations as it has a low discrimination against O_2 (Badger and Bek, 2008; Singer *et al*., 2011). Thus, consistent with this finding and a prior survey that reported a high incidence of form II RuBisCO genes in the system (Emerson *et al*., 2016), we conclude that organisms encoding this form of RuBisCO and the CBB cycle for carbon fixation have a selective advantage under high CO_2 and low (or non-existent) O_2 concentrations.

Fifteen of the 150 genomes encode genes necessary to fix CO_2 by the CBB cycle (40.8% by relative abundance) and 22 by the WL pathway (32.8% by relative abundance). These fractions sum to >70% by relative abundance because two organisms, members of the Cyanobacteria and Rhodobacteraceae respectively, have both pathways (Fig. 5). This is unusual, and has been interpreted as an adaptation to varying CO_2 concentrations (Markert *et al*., 2007; Berg, 2011) enabling these organisms greater metabolic flexibility in the ecosystem. The reverse citric acid cycle, with the key enzymes ATP-citrate lyase and the 2-oxoglutarate synthase, was present in only two organisms (0.17% by relative abundance). Even when analyzing the entire unbinned metagenome we identified only five ATP-citrate lyases (three in binned genomes, one unbinned on a scaffold belonging to a *Sulfurimonas* and one on an unknown scaffold).

A few genes annotated as ATP-citrate lyase are present in some publicly available CPR genomes. Like the sequence from *Ca*. ‘W. wanneri’, their phylogenetic placement (Supporting Information Fig. S-9) is uncertain. Their

sequences do not strongly group with sequences from autotrophs and are distant to those found in Eukaryotes, where this enzyme participates in a different metabolic cycle. Given that neither a complete TCA cycle nor r-TCA has been identified in CPR, these enzymes may metabolize environmentally derived citrate. This observation underlines the riskiness of inferring autotrophic capabilities based on the detection of ATP-citrate lyase genes via PCR amplification (Alfreider and Vogt, 2012), or in unbinned/unassembled metagenomic data.

The three carbon fixation pathways detected in the system differ substantially in the energy required to build one pyruvate from bicarbonate and/or CO_2 (Berg *et al*., 2010). The WL pathway is the most energy efficient pathway (~1 ATP per pyruvate) and may contribute to the dominance of '*Ca*. Altiarchaeum' in the ecosystem. Based on ATP cost alone, the high frequency of the CBB cycle, despite its high cost (~7 ATP per pyruvate), and the low frequency of the reverse citric acid cycle, despite its low cost (~2 ATP per pyruvate; in two organisms only), would seem to be paradoxical. An explanation may be found in the different carbon species used for fixation in the different pathways (CBB and WL pathway CO_2 only, r-TCA CO_2 and bicarbonate). The need for transporters and anhydrases to maintain equilibrium of CO_2 and bicarbonate may make the r-TCA cycle less energy favorable than the CBB cycle in this high CO_2 environment. A more plausible explanation may be periodic access to O_2. Almost all of the genomes for organisms with CBB cycle also encoded for cytochrome *c*/ubiquinol oxidase (complex IV, Fig. 5), which would provide a selective advantage over anaerobes with the r-TCA cycle. So far, O_2 has only been detected at trace levels in the subsurface aquifer fluids, and is likely from atmospheric contamination (Mayo *et al*., 1991; Heath *et al*., 2009). However, the O_2 concentration may be sufficient to give organisms with the CBB cycle an energetic advantage over those with the r-TCA cycle yet low enough enabling them to fix CO_2 with form II RuBisCO.

The great diversity and abundance of carbon fixers suggests that the Crystal Geyser harbours many environmental niches in which different autotrophs grow. This is highlighted by the fact that two organisms' genomes encode for two carbon fixation pathways, the energy efficient but O_2 sensitive WL pathway and the CBB cycle, enabling growth under different geochemical conditions. Further, the detection of two Cyanobacteria with genes for photosynthesis indicates the presence of organisms living in surface fluids. The recovered genomes of 150 different bacterial and archaeal species represent an important foundation for future analyses, which will investigate how these organisms respond to varying geochemical conditions throughout the geyser's multi-day eruption cycle.

Conclusions

Subsurface microbial life remains little explored with regard to phylogenetic and metabolic diversity. Further, little is known about adaptations to the

variety of environment types that exist in the subsurface and the resources that sustain organisms that reside in them. Genomic resolution of microbial communities can provide insight into the metabolic capacities of uncultivated organism, and thus begin to address these questions. Using a novel approach, we reconstructed genomes for the majority of relatively abundant species in the system, a result that underlines the power of metagenomics for ecosystem description. One highlight of this approach was the recovery of two novel phylum-level lineages, one closely associated with the CPR (*Ca*. 'Wirthbacteria') and another indigenous to various anoxic subsurface ecosystems (*Ca*. 'Desantisbacteria'). The methods presented in this communication should improve binning efforts of future metagenomic studies.

The carbon cycle in the Crystal Geyser aquifer includes conversion of inorganic to organic carbon and the recycling of short-chain carboxylic acids, probably derived from fermentative organisms (Fig. 6). The capacity for mixotrophy confers metabolic flexibility to some organisms. Autotrophs also likely benefit from H_2 produced by the fermenters. Many of the primary producers harness energy from oxidation and reduction of various inorganic compounds. Their biochemical pathways that incorporate CO_2 rather than bicarbonate and the presence of two carbon fixation pathways in one genome may represent environmental adaptations. Overall, the findings presented in this communication provide new insight into the diversity and metabolism of subsurface microbial life under extreme conditions of CO_2 saturation.

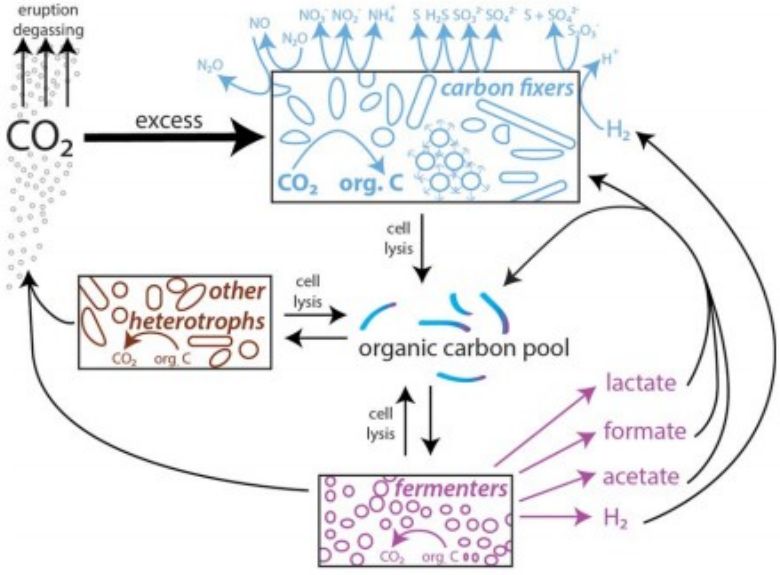

Fig. 6. Metabolic model of the microbial community found at Crystal Geyser, which is based on CO_2 as initial carbon source. Note the recycling of H_2 and short-chain carboxylic acids by autotrophs, which are thus predicted to function as mixotrophs.

Experimental procedures

Sampling and DNA extraction

Samples were taken from the Crystal Geyser system in Utah (USA), as described previously (Probst *et al*., 2014; Emerson *et al*., 2016). Sixty-five

litres of water were filtered through nine 3.0-µm filters resulting in approximately 7 l of water filtrate per filter. DNA was extracted from eight of these filters (named CG1 A-H herein) using the PowerSoil DNA extraction kit (MoBio, Carlsbad, USA), and four of the eight samples underwent freezing in liquid nitrogen and thawing at 65°C prior to DNA extraction. Metagenomic DNA of these samples was used for library preparation and sequencing as described previously (Emerson *et al*., 2016). This shallow sequencing data were used for series binning (Sharon *et al*., 2013) of samples CG1_0.2 and CG2_3.0, which have been previously acquired but not used for genome reconstruction (Emerson *et al*., 2016). Sample CG2_3.0 and CG1_0.2 were collected from geyser water that was sampled 32 hrs after the first eruption by sequential filtration onto a 3.0-µm and a 0.2-µm filter respectively. An overview of the samples and corresponding statistics is given in Supporting Information Table S1.

Metagenomic datasets and assembly

Reads were quality filtered (SICKLE Version 1.21, https://github.com/najoshi/sickle, default parameters) and assembled as well as scaffolded by IBDA_UD (Peng *et al*., 2012). Scaffolding errors were corrected using MISS (Sharon *et al*., unpublished), a tool that searches and fixes gaps in the assembly based on mapped reads that exhibit inconsistencies between raw reads and assembly. The two main samples CG1_0.2 and CG2_3.0 used for binning in this study resulted in 648 and 713 Mbps of assembled scaffolds respectively. Gene prediction of metagenomes was performed using Prodigal (meta function) (Hyatt *et al*., 2010). Sequencing data is publicly available under the BioProjects PRJNA229517 and PRJNA297582.

Tetranucleotide frequency-based binning, differential coverage binning, subsampling of metagenomic libraries for binning, and genome completeness estimation are described in the Supporting Information Experimental Procedures. Overall, ten genomes could be improved regarding their completeness via binning of a subsampled assembly.

Integrating different binning methods to improve recovery of near-complete genomes

Different binning methods were employed in order to retrieve a representative number of genomes of organisms for the ecosystem. We utilized tetranucleotide frequency and series-based ESOM binning, and series-based ABAWACA binning (see above), whereas this ordering reflects the decreasing credibility. The sequential binning approach was individually applied to the CG1_0.2 and CG2_3.0 samples; an overview of this approach is provided in Supporting Information Fig. S-1.

Using ggKbase (ggkbase.berkeley.edu) bins generated from tetranucleotide frequency clustering were improved using information from number of (multiple) single copy genes, %GC content, sequencing coverage, and

taxonomic information (removing or recruiting scaffolds). Improved bins were then evaluated using series binning of the original assembly by first finding the best matching bin (at least 50% overlap of nucleotides calculated from the shared scaffolds) and then evaluating an increase or decrease in the estimated genome completeness. Genomes with higher completeness were favoured over genomes of lower completeness, taking the number of duplicated single copy genes into account. Bins not previously included, but with at least 95% sequence match in the unbinned fraction from the tetranucleotide frequency binning were also selected based on bacterial single copy gene inventory. The resulting bins were then further improved with the same methods using bins derived from automated ABAWACA binning of series data.

The resulting bins were then again improved in ggKbase regarding consistency in GC, coverage and phylogenetic profile of each scaffold for each genomic bin. Only bins representative of one single genome and at least 70% complete were considered for further analyses are publicly available under BioProject PRJNA297582. Bins with detectable contamination, i.e. more than three multiple single-copy genes or contamination based on taxonomic identity of scaffolds were not included in this study.

Dereplication of genomic bins

From CG1_0.2, CG2_3.0 and their subassemblies, 227 genomes with a completeness of at least 70% were generated. These genomes were de-replicated and only the most complete representative genome of each organism was used for further analysis. For de-replication, we first generated an alignment of all scaffolds within one bin individually against scaffolds of all other bins using NUCmer (Delcher *et al*., 2002) at 98% nucleotide level or greater and recorded its percent identity across all bins. The best representative with a similarity of >70% (i.e. the genomes shared >70% of the nucleotides at 98% similarity level in alignments) was chosen based on the number of archaeal/bacterial single copy genes, multiple single copy genes and genome length, favouring high number of single copy genes, low number of multiple single copy genes and high genome length.

Automated genome curation, genome abundance estimation, phylogenomic analyses, and metabolic overview (Supporting Information Fig. S-5C) are presented in the Supporting Information Experimental and Procedures.

Genome annotation

For each curated scaffold, protein-coding genes were predicted using Prodigal (Hyatt *et al*., 2010). Ribosomal RNA genes were searched for using Rfam (Nawrocki *et al*., 2015); specifically, 16S rRNA genes were identified using SSU-Align (Nawrocki, 2009). Predicted proteins were annotated using USEARCH (ublast [Edgar, 2010]) against UniProt (UniProt, 2010), Uniref90 (Suzek *et al*., 2007), KEGG (Kanehisa and Goto, 2000) and a custom in-house database (Brown *et al*., 2015) by determining the reverse best blast hit.

These annotations are also stored on ggKbase (Supporting Information Table S-1) (Wrighton *et al*., 2012), where they are updated as new annotations are added or changed.

Reconstruction of pathways for autotrophy

Autotrophic pathways within each reconstructed genome were called if all key enzymes (Berg *et al*., 2010) of one of the six known pathways and at least 60% of the entire pathway were present (known autotrophs in the system like '*Ca*. Altiarchaeum' (Probst *et al*., 2014) or Cyanobacteria showed at least 60% pathway coverage). For certain key enzymes further criteria were required. For the CO-DH/ACS complex, the key enzyme for the WL pathway, at least three different subunits had to be present. For ATP-citrate lyase, a key enzyme for the r-TCA cycle, both subunits and the active centre (GHAGA) (Kanao *et al*., 2001) were necessary. The alpha subunit was further classified via phylogeny; related sequences were retrieved from NCBI via word search, filtered for the active center and served as a reference set. Sequences from this study and other groundwater systems (unpublished) were aligned (Edgar, 2004) with the reference set, end-trimmed, and used for reconstructing a phylogenetic tree (model = PROTGAMMALG, 100 bootstrap replications) (Stamatakis *et al*., 2005). RuBisCO forms were identified using the catalytic subunit as described previously (Wrighton *et al*., 2012).

Reconstruction of other metabolic processes

Other metabolic processes mentioned in the paper were inferred from the presence of corresponding enzymes annotated in the genomes (e.g. conversion of CO_2 to bicarbonate via carbonic anhydrase). Processes that necessitate more than one enzymatic reaction are inferred from the presence of all corresponding enzymes and listed in Supporting Information Table S-4. Ublast (Edgar, 2010) against the entire predicted proteome was used to find candidates for the dimethyl sulfoxide (DMSO) reductase superfamily against a previously established database (e-value cutoff 1e-2) (Castelle *et al*., 2013). Combined with genes annotated as members of this superfamily, a phylogenetic classification was performed (e.g. nitrate reductases or nitrite oxidoreductases) (Castelle *et al*., 2013). Calling potential nitrogen fixation via the NifHDK complex required the presence of at least two different subunits encoded in the genome. Metabolic predictions for the Ca. 'W. wanneri' were confirmed using the MaGe platform for genome annotation (Vallenet *et al*., 2013). Comparative genomics for members of the Ca. 'Desantisbacteria' was performed based on shared KEGG orthologies (KOs) as follows. Predicted proteins of each genome were searched against a database of HMMs representing all the KOs (Kanehisa and Goto, 2000). The HMM database was compiled using the HMMER suite (Finn *et al*., 2011), based on assignment of proteins to KOs according to KEGG FTP Release 2015-06-22. Individual trusted treshold were calculated by running HMM search of all the proteins with assigned KOs against the HMM database.

Acknowledgements

JFB was supported as part of the Center for Nanoscale Controls on Geologic CO_2 (NCGC), an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under Award # DE-AC02-05CH11231. AJP was supported by the DFG grant PR 1603/1-1. We thank M. Cathryn Ryan and Bethany Ladd for scientific discussions about the Crystal Geyser system, particularly its hydrogeology and geochemistry. We thank Alex Hernsdorf for providing genomes of lineage ACD39 to improve phylogenetic reconstruction.

References

Alfreider, A., and Vogt, C. (2012) Genetic evidence for bacterial chemolithoautotrophy based on the reductive tricarboxylic acid cycle in groundwater systems. *Microbes Environ* 27: 209– 214.

Anantharaman, K., Brown, C.T., Burstein, D., Castelle, C.J., Probst, A.J., Thomas, B.C., *et al*. (2016) Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum. *PeerJ* 4: e1607.

Badger, M.R., and Bek, E.J. (2008) Multiple Rubisco forms in proteobacteria: their functional significance in relation to CO_2 acquisition by the CBB cycle. *Journal of Experimanetal Botany* 59: 1525– 1541.

Baer, J.L., and Rigby, J.K. (1978) Geology of the Crystal Geyser and environmental implications of its effluent, Grand County, Utah. *Utah Geol* 5: 125– 130.

Baker, B.J., Tyson, G.W., Webb, R.I., Flanagan, J., Hugenholtz, P., Allen, E.E., and Banfield, J.F. (2006) Lineages of acidophilic archaea revealed by community genomic analysis. *Science* 314: 1933– 1935.

Berg, I.A. (2011) Ecological aspects of the distribution of different autotrophic CO_2 fixation pathways. *Appl Environ Microbiol* 77: 1925– 1936.

Berg, I.A., Kockelkorn, D., Ramos-Vera, W.H., Say, R.F., Zarzycki, J., Hügler, M., *et al*. (2010) Autotrophic carbon fixation in archaea. *Nat Rev Microbiol* 8: 447– 460.

Bickle, M.J. (2009) Geological carbon storage. *Nat Geosci* 2: 815– 818.

Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., *et al*. (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 523: 208– 211.

Bryan, T.S. (2008) *The geysers of Yellowstone*. Fourth edition, University Press of Colorado.

Castelle, C.J., Hug, L.A., Wrighton, K.C., Thomas, B.C., Williams, K.H., Wu, D., *et al*. (2013) Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nat Commun* 4: 2120.

Castelle, C.J., Wrighton, K.C., Thomas, B.C., Hug, L.A., Brown, C.T., Wilkins, M.J., *et al*. (2015) Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling. *Curr Biol.* 25: 690– 701.

Delcher, A.L., Phillippy, A., Carlton, J., and Salzberg, S.L. (2002) Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* 30: 2478– 2483.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids Res* 32: 1792– 1797.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460– 2461.

Emerson, J.B., Thomas, B.C., Alvarez, W., and Banfield, J.F. (2016) Metagenomic analysis of a high carbon dioxide subsurface microbial community populated by chemolithoautotrophs and bacteria and archaea from candidate phyla. *Environ Microbiol* 18: 1686– 1703.

Evans, J.P., Heath, J., Shipton, Z.K., Kolesar, P.T., Dockrill, B., Williams, A., *et al*. (2004) Natural leaking CO_2-charged systems as analogs for geologic sequestration sites. In *Third Annual Conference on Carbon Capture and Sequestration, Alexandria, VA*.

Finn, R.D., Clements, J., and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 39: W29– W37.

Friedmann, S.J. (2007) Geological carbon dioxide sequestration. *Elements* 3: 179– 184.

Glöckner, J., Kube, M., Shrestha, P.M., Weber, M., Glöckner, F.O., Reinhardt, R., and Liesack, W. (2010) Phylogenetic diversity and metagenomics of candidate division OP3. *Environ Microbiol* 12: 1218– 1229.

Gong, J., Qing, Y., Guo, X., and Warren, A. (2014) “ Candidatus Sonnebornia yantaiensis”, a member of candidate division OD1, as intracellular bacteria of the ciliated protist Paramecium bursaria (Ciliophora, Oligohymenophorea). *Syst Appl Microbiol* 37: 35– 41.

Gouveia, F.J., and Friedmann, S.J. (2006) *Timing and prediction of CO2 eruptions from Crystal Geyser, UT*, United States: Department of Energy.

Greening, C., Biswas, A., Carere, C.R., Jackson, C.J., Taylor, M.C., Stott, M.B., *et al*. (2015) Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival. *ISME J.* 10: 761– 777.

Han, W.S., Lu, M., McPherson, B.J., Keating, E.H., Moore, J., Park, E., *et al*. (2013) Characteristics of CO_2-driven cold-water geyser, Crystal Geyser in Utah: experimental observation and mechanism analyses. *Geofluids* 13: 283– 297.

Handley, K.M., VerBerkmoes, N.C., Steefel, C.I., Williams, K.H., Sharon, I., Miller, C.S., *et al*. (2013) Biostimulation induces syntrophic interactions that impact C, S and N cycling in a sediment microbial community. *ISME J* 7: 800– 816.

He, X., McLean, J.S., Edlund, A., Yooseph, S., Hall, A.P., Liu, S.Y., *et al*. (2015) Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proc Natl Acad Sci USA* 112: 244– 249.

Heath, J.E., Lachmar, T.E., Evans, J.P., Kolesar, P.T., and Williams, A.P. (2009) Hydrogeochemical characterization of leaking, carbon dioxide-charged fault zones in east-central Utah, with implications for geologic carbon storage. *Carbon Sequestration and Its Role in the Global Carbon Cycle* 183: 147– 158.

Herrmann, M., Rusznyák, A., Akob, D.M., Schulze, I., Opitz, S., Totsche, K.U., and Kusel, K. (2015) Large fractions of CO_2-fixing microorganisms in pristine limestone aquifers appear to be involved in the oxidation of reduced sulfur and nitrogen compounds. *Appl Environ Microbiol* 81: 2384– 2394.

Hu, P., Tom, L., Singh, A., Thomas, B.C., Baker, B.J., Piceno, Y.M., *et al*. (2016) Genome-resolved metagenomic analysis reveals roles for candidate phyla and other microbial community members in biogeochemical transformations in oil reservoirs. *MBio* 7: Article e01669-15.

Hug, L.A., Castelle, C.J., Wrighton, K.C., Thomas, B.C., Sharon, I., Frischkorn, K.R., *et al*. (2013) Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome* 1: 22.

Hugenholtz, P., Pitulle, C., Hershberger, K.L., and Pace, N.R. (1998) Novel division level bacterial diversity in a Yellowstone hot spring. *J Bacteriol* 180: 366– 376.

Hyatt, D., Chen, G.L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform* 11: 119.

Kanao, T., Fukui, T., Atomi, H., and Imanaka, T. (2001) ATP-citrate lyase from the green sulfur bacterium Chlorobium limicola is a heteromeric enzyme composed of two distinct gene products. *Eur J Biochem* 268: 1670– 1678.

Kanehisa, M., and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28: 27– 30.

Kantor, R.S., Wrighton, K.C., Handley, K.M., Sharon, I., Hug, L.A., Castelle, C.J., *et al*. (2013) Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. *MBio* 4: e00708– e00713.

Kolinko, S., Richter, M., Glöckner, F.O., Brachmann, A., and Schüler, D. (2015) Single-cell genomics of uncultivated deep-branching magnetotactic bacteria reveals a conserved set of magnetosome genes. *Environ Microbiol* 18: 21– 37.

Kolinko, S., Jogler, C., Katzmann, E., Wanner, G., Peplies, J., and Schüler, D. (2012) Single-cell analysis reveals a novel uncultivated magnetotactic bacterium within the candidate division OP3. *Environ Microbiol* 14: 1709–1721.

Könneke, M., Schubert, D.M., Brown, P.C., Hügler, M., Standfest, S., Schwander, T., *et al*. (2014) Ammonia-oxidizing archaea use the most energy-efficient aerobic pathway for CO_2 fixation. *Proc Natl Acad Sci USA* 111: 8239– 8244.

Lewicki, J.L., Birkholzer, J., and Tsang, C.-F. (2007) Natural and industrial analogues for leakage of CO_2 from storage reservoirs: identification of features, events, and processes and lessons learned. *Environ Geol* 52: 457–467.

Markert, S., Arndt, C., Felbeck, H., Becher, D., Sievert, S.M., Hugler, M., *et al*. (2007) Physiological proteomics of the uncultured endosymbiont of Riftia pachyptila. *Science* 315: 247– 250.

Marreiros, B.C., Batista, A.P., Duarte, A.M., and Pereira, M.M. (2013) A missing link between complex I and group 4 membrane-bound [NiFe] hydrogenases. *Biochim Biophys Acta* 1827: 198– 209.

Mayo, A.L., Shrum, D.B., and Chidsey Jr, T.C. (1991) *Factors Contributing to Exsolving Carbon Dioxide in Ground Water Systems in the Colorado Plateau, Utah*, pp. 335–342.

Mu, A., Boreham, C., Leong, H.X., Haese, R.R., and Moreau, J.W. (2014) Changes in the deep subsurface microbial biosphere resulting from a field-scale CO_2 geosequestration experiment. *Front Microbiol* 5: 209.

Murray, C. (1989) The cold water geyser of Utah, II: observation of Crystal Geyser. *Geyser Observation Study Assoc* 2: 133– 139.

Nawrocki, E.P. (2009) *Structural RNA Homology Search and Alignment Using Covariance Models*, Washington University in St. Louis.

Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., *et al*. (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res* 43: D130– D137.

Nunoura, T., Takaki, Y., Kakuta, J., Nishi, S., Sugahara, J., Kazama, H., *et al*. (2011) Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. *Nucleic Acids Res* 39: 3204– 3223.

Peng, Y., Leung, H.C., Yiu, S.M., and Chin, F.Y. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 28: 1420– 1428.

Porter, R.T.J., Fairweather, M., Pourkashanian, M., and Woolley, R.M. (2015) The range and level of impurities in CO_2 streams from different carbon capture sources. *Int J Greenhouse Gas Control* 36: 161– 174.

Probst, A.J., Weinmaier, T., Raymann, K., Perras, A., Emerson, J.B., Rattei, T., *et al*. (2014) Biology of a widespread uncultivated archaeon that contributes to carbon fixation in the subsurface. *Nat Commun* 5: 5497.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., and Glöckner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35: 7188– 7196.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., *et al*. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* 499: 431– 437.

Santillan, E.F.U., Shanahan, T.M., Omelon, C.R., Major, J.R., and Bennett, P.C. (2015) Isolation and characterization of a CO_2-tolerant Lactobacillus strain from Crystal Geyser, Utah, USA. *Front Earth Sci* 3: 41.

Sato, T., Atomi, H., and Imanaka, T. (2007) Archaeal type III RuBisCOs function in a pathway for AMP metabolism. *Science* 315: 1003– 1006.

Schut, G.J., and Adams, M.W. (2009) The iron-hydrogenase of Thermotoga maritima utilizes ferredoxin and NADH synergistically: a new perspective on anaerobic hydrogen production. *J Bacteriol* 191: 4451– 4457.

Sharon, I., Morowitz, M.J., Thomas, B.C., Costello, E.K., Relman, D.A., and Banfield, J.F. (2013) Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* 23: 111– 120.

Shipton, Z.K., Evans, J.P., Dockrill, B., Heath, J., Williams, A., Kirchner, D., and Kolesar, P.T. (2006) Natural leaking CO_2-charged systems as analogs for failed geologic storage reservoirs. In *Carbon Dioxide Capture for Storage in Deep Geologic Formations–Results from the CO2 Capture Project*. D.C. Thomas, and S.M. Benson (eds): Elsevier, pp. 699– 712.

Sieber, J.R., McInerney, M.J., and Gunsalus, R.P. (2012) Genomic insights into syntrophy: the paradigm for anaerobic metabolic cooperation. *Annu Rev Microbiol* 66: 429– 452.

Singer, E., Emerson, D., Webb, E.A., Barco, R.A., Kuenen, J.G., Nelson, W.C., *et al*. (2011) Mariprofundus ferrooxydans PV-1 the first genome of a marine Fe(II) oxidizing Zetaproteobacterium. *PLoS One* 6: e25386.

Soro, V., Dutton, L.C., Sprague, S.V., Nobbs, A.H., Ireland, A.J., Sandy, J.R., *et al*. (2014) Axenic culture of a candidate division TM7 bacterium from the human oral cavity and biofilm interactions with other oral bacteria. *Appl Environ Microbiol* 80: 6480– 6489.

Stamatakis, A., Ludwig, T., and Meier, H. (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics* 21: 456– 463.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 23: 1282– 1288.

UniProt, C. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* 38: D142– D148.

Vallenet, D., Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., *et al*. (2013) MicroScope - an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res* 41: D636– D647.

Vignais, P.M., and Colbeau, A. (2004) Molecular biology of microbial hydrogenases. *Current Issues Mol Biol* 6: 159– 188.

Wandrey, M., Pellizari, L., Zettlitzer, M., and Würdemann, H. (2011a) Microbial community and inorganic fluid analysis during CO_2 storage within the frame of CO 2 SINK–Long-term experiments under in situ conditions. *Energy Procedia* 4: 3651– 3657.

Wandrey, M., Fischer, S., Zemke, K., Liebscher, A., Scherf, A.-K., Vieth-Hillebrand, A., *et al*. (2011b) Monitoring petrophysical, mineralogical, geochemical and microbiological effects of CO_2 exposure—Results of long-term experiments under in situ conditions. *Energy Procedia* 4: 3644– 3650.

Watson, Z.T., Han, W.S., Keating, E.H., Jungm, N.-H., and Lu, M. (2014) Eruption dynamics of CO_2-driven cold-water geysers: Crystal, Tenmile geysers in Utah and Chimayó geyser in New Mexico. *Earth Planetary Sci Lett* 408: 272– 284.

White, C.M., Strazisar, B.R., Granite, E.J., Hoffman, J.S., Pennline, H.W., and Air and Waste Management Association (2003) Separation and capture of CO_2 from large stationary sources and sequestration in geological formations–coalbeds and deep saline aquifers. *J Air Waste Manag Assoc* 53: 645– 715.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* 95: 6578– 6583.

Wilson, E.J., Johnson, T.L., and Keith, D.W. (2003) Regulating the ultimate sink: managing the risks of geologic CO_2 storage. *Environ Sci Technol* 37: 3476– 3483.

Wrighton, K.C., Thomas, B.C., Sharon, I., Miller, C.S., Castelle, C.J., VerBerkmoes, N.C., *et al*. (2012) Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* 337: 1661–1665.

Wrighton, K.C., Castelle, C.J., Wilkins, M.J., Hug, L.A., Sharon, I., Thomas, B.C., *et al*. (2014) Metabolic interdependencies between phylogenetically novel fermenters and respiratory organisms in an unconfined aquifer. *ISME J* 8: 1452– 1463.

Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F.O., Ludwig, W., Schleifer, K.H., *et al*. (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* 12: 635–645.